Aral de Moor

Amsterdam, Netherlands · aral.demoor+cv@gmail.com · +31 6 5196 5297
github.com/Ar4l · huggingface.co/Ar4l · linkedin.com/in/aral-dm · aral.cc
EU citizen.

ML Engineer building the evaluation infrastructure behind Mellum, JetBrains' open-weights agentic coding model — benchmarks, sandboxed execution, and the statistics that make results trustworthy. Ran its supervised fine-tuning and architectural ablations; first author of an ACM SIGSOFT Distinguished Paper.

Experience

ML Engineer, Mellum Team — JetBrains 2025 – Present

- Built and own Mellum's evaluation suite (12B-A2.5B mixture-of-experts, 128K context): vLLM/llama-cpp serving, sandboxing, LLM-as-a-judge scoring, and hypothesis testing, on ZenML/Kubernetes.
- Participated in Mellum 2's SFT end to end on 128×H200: 47B Instruct, 167B Thinking tokens using Megatron-Bridge, expert parallelism 8, context parallelism 8.

ML Engineer, AI for Code Team — JetBrains 2024 – 2025

- Owned trigger models end to end — large data processing, custom evaluation, IDE integration, online A/B: on-device CatBoost classifiers over ∼120 telemetry features gating inline completion. Shipped on by default for every language across all JetBrains IDEs, reaching millions of developers and saving 20–30% of code-completion inference.
- Mentored the engineer who took the model from Kotlin-only to polyglot; first author of the IDE'26 @ ICSE paper.

Scientific Developer, AISE Lab — TU Delft & JetBrains 2023 – 2024

- Filtered 34% of completion invocations at a 3% false-negative rate over 200K interactions, using a PyTorch transformer fusing code context and IDE telemetry, validated live with 34 developers.
- Supervised five BSc theses on sample-efficient small-transformer pre-training: activation functions, sparse feed-forward layers, tokenisation, grouped-query attention dimensionality, Infini-Attention.

Publications

- **Mellum2 Technical Report** — co-author. arxiv.org/abs/2605.31268; weights on Hugging Face, Apache 2.0, ∼70K downloads/month across 16 repositories (Aug 2026).
- **A Transformer-Based Approach for Smart Invocation of Automatic Code Completion** — first author. AIware 2024 @ FSE; *ACM SIGSOFT Distinguished Paper Award*. arxiv.org/abs/2405.14753
- **Control Models for In-IDE Code Completion** — first author. IDE'26 @ ICSE 2026. arxiv.org/abs/2601.20223
- **BSc thesis** — compressed CodeGPT 15× via layer reduction and 1-bit-weight / 8-bit-activation quantisation through knowledge distillation, at 84% of original accuracy and ∼2× inference speed. TU Delft repository

Education

BSc Computer Science & Engineering, Data Science track — Delft University of Technology 2020 – 2023
Graduated June 2023 with 20 EC of additional coursework; research thesis on LLM compression.

Skills

- **Evaluation:** benchmark harness design and implementation, sandboxed code execution, LLM-as-a-judge, bootstrap and confidence intervals, cluster-robust significance testing, online A/B evaluation.
- **Post-training:** supervised fine-tuning (SFT), knowledge distillation, on-policy distillation.
- **Infrastructure:** PyTorch, vLLM, Hugging Face stack, Megatron-LM / Megatron-Bridge, ZenML, Kubernetes, Docker, CatBoost, HPC GPU clusters.
- **Programming:** Python, shell, Rust, Kotlin/Java. **Languages:** English, Dutch, Turkish